論 文 の 要 旨

題 目 **A study of depth estimation and depth based applications**
(奥行きデータの推定とその応用に関する研究)

氏 名 TRAN ANH TU

Three dimensional video ($3DV$) and multi-view imaging technologies may be the next step in the evolution of motion picture formats, as we presently witness the appearance of $3D$ displays, multi-camera systems with dense or sparse camera configuration, coding systems. Going with the demand of entertainment and progressive development of digital devices, developing $3D$ processing algorithms, related applications and systems have been attracted extensive attentions in the industrial and research communities. Depth inference from stereo and multi-view images is one of the most fundamental techniques in $3D$ digital imaging applications since it provides the perception and visualization of the real word environment in $3DV$ as well as a useful cue for other applications.

This thesis devotes to firstly study depth estimation from multi-view images and then use this useful information for three applications including: one of the key applications in $3DTV$, namely free viewpoint synthesis, and other two applications object segmentation and multiple moving object tracking.

The first part (Chapter 3) of this thesis addresses the problem of depth estimation from multiple views. The depth information disappears after taking an $2D$ image from a $3D$ scene. To recover this missing information, the depth can be estimated from two or more images by finding the correspondence pairs among them. Initially, we introduce the basic geometric model that enables the triangulation of corresponding pixel points across the views. While the basic physics and geometry relating visual disparity to scene structure are well understood automatically measuring this disparity by establishing dense and accurate inter image correspondence is a challenging task. Some difficulties such as the unable setting the identical internal cameras' parameters, the change of illumination across the views, texture-less regions and occlusion can result in an unreliable identification of the point-correspondences and thus in inaccurate depth values. Next, we review the previous works on estimation of depth image using a single image/mono video, two views and multiple views. Finally, we have proposed a method that allows the use of several un-rectified images simultaneously to estimate a consistency and reliability depth image. We have introduced three constraints, i.e. intra-line, inter-line and inter-view smoothness constraint, which enforce smooth variations of depth value in the scanline, across scanline and consistent depth value across the views. The proposed algorithm combines two stages: the first stage serves as a calculation of initial depth images and the second stage enhances the depth initial depth images in the first step by enforcing consistent depth across the views. The three smooth constraints can be efficiently integrated into one dimensional optimization dynamic program algorithm. Experiments have shown that the proposed method yields reasonably depth image quality for various multi-view data sets.

After investigating and presenting the depth estimation algorithm, the next part (Chapter 4) of this thesis focuses on the depth based image rendering for $3D$ video and $3DTV$ systems. In $3DV$ / $3DTV$, the viewer can ideally navigate through the $3D$ domain and selects his own viewpoint. The chosen viewpoint may not only be selected from available multi-view camera views, but also any viewpoint between these cameras. Obviously, this feature requires a smart synthesis algorithm that allows free-viewpoint view rendering. In chapter 4, we have reviewed the recent advancements in viewpoint synthesis for $3DTV$ and then proposed a novel method and showed its performance. Our contribution is a novel synthesis method that enables to render a free-viewpoint from multiple existing cameras. The proposed method solves the main problems of depth based synthesis by applying forward depth map following with inverse warping texture, performing pixel classification to generate an initial new view from stable pixels and using Graph cut to select the best candidate for unstable pixels. By defining the types of pixels and using Graph cuts, the color is consistent and the pixels wrapped incorrectly because of inaccuracy depth maps are removed. The remained disoccluded pixels are inpainted by using depth and texture

neighboring pixel value. Considering depth information for inpainting, blurring between foreground and background textures are reduced. Experimental results show that the proposed method has strength in artifact reduction. Objective evaluation has shown that our method get a significant gain in Peak Signal Noise Ratio (*PSNR*) and Structure Similarity Index (*SSIM*) comparing to some other existing methods. Another advantage of our method is that we can use a set of un-rectified images in multi-view system to create a new view with higher quality

As the estimated depth information available, our concern is to apply the usefully estimated $3D$ information for the object segmentation method (Chapter 5). Even though image segmentation has been addressed in extensive literatures, the results are not satisfactory in many situations. A major difficulty lies in the fact that semantic objects are not homogeneous with respect to the low-level features in single image, such as color or texture properties. Fortunately, depth information recovered from multi-view serves as an important cue for segmentation. We have proposed a method using both depth and color cues, which requires no interactive operation, to segment human object from multi-view video. Our method consist of two stages: for initial frame of the video sequence, the interested object is automatically extracted based on saliency model and iterated Graph cut. After having segmented object in first frame, from the second frame we have proposed a method combining Bayesian estimation and minimizing energy function using Graph cut to segment object. We use Gaussian Mixture Model (*GMM*) in RGB space for the color cue and histogram model for depth cue. Based on these probabilistic models, the probability of each pixel to be in foreground is computed base on Bayesian estimation and the results are used to create the tri-map including foreground (F), background (B) and uncertain region (U). Graph cut is then performed on the uncertain region. In the energy function for Graph cut optimization, the color, depth and spatial-temporal coherence are integrated in data term and the penalty cost of the neighboring pixels with different labels is encoded in smoothness term. Experiment results on test sequences are encouraging and showed that our method is more effective than the case using only color cue.

The final work in this thesis is to using the estimated depth information for object tracking (Chapter 6). Detection and tracking of objects is very importance research area of computer vision and has a wide range of application. Many researchers have investigated object tracking and different approaches have been presented. Some of approaches can achieve good results in some cases, such as when the target object has distinct color distribution from the background. However, multi objects tracking is still a difficult task due to various aspects, including inaccurate motion vector estimation, variation of the non-rigid object appearance and confusions in multiple targets' identities when their projections in the image are close. Moreover, object regions with various tracking issues such as appearance and disappearance, splitting and merging, without and with occlusion should be dealt with in the tracking algorithm. We have proposed a novel tracking method aiming at detecting objects and maintaining their label/identification over the time. The main key factors of this method are to use depth information and different strategies to track objects under various occlusion scenarios. The foreground objects are detected and refined by background subtraction and shadow cancellation. The occlusion detection is based on information of foreground blobs in successive frames. The occlusion regions are projected to the projection plane XZ (ground plane) to analysis occlusion situations. According to the occlusion analysis results, different objects correspondence strategies are introduced to track object under various occlusion scenarios including tracking occluded objects in similar depth layer and in different depth layers. The experimental results show that our proposed method can track the moving objects under the most typical and challenging occlusion scenarios.